# The Role of Administrative Data in the Big Data Revolution in Social Science Research

Social Science Research xxx (2016) 1–12

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

# The role of administrative data in the big data revolution in social science research

Roxanne Connelly [a, *], Christopher J. Playford [b], Vernon Gayle [c], Chris Dibben [d]

[a] Department of Sociology, University of Warwick, Social Sciences Building, The University of Warwick, Coventry, CV4 7AL, UK
[b] Administrative Data Research Centre – Scotland, University of Edinburgh, 9 Edinburgh Bioquarter, Little France Road, Edinburgh, EH16 4UX, UK
[c] School of Social and Political Science, University of Edinburgh, 18 Buccleuch Place, Edinburgh, EH8 9LN, UK
[d] School of Geosciences, University of Edinburgh, Geography Building, Drummond Street, Edinburgh, EH8 9XP, UK

ARTICLE INFO

*Article history:*
Received 8 July 2015
Received in revised form 5 April 2016
Accepted 13 April 2016
Available online xxx

*Keywords:*
Big data
Administrative data
Data management
Data quality
Data access

ABSTRACT

The term big data is currently a buzzword in social science, however its precise meaning is ambiguous. In this paper we focus on administrative data which is a distinctive form of big data. Exciting new opportunities for social science research will be afforded by new administrative data resources, but these are currently under appreciated by the research community. The central aim of this paper is to discuss the challenges associated with administrative data. We emphasise that it is critical for researchers to carefully consider how administrative data has been produced. We conclude that administrative datasets have the potential to contribute to the development of high-quality and impactful social science research, and should not be overlooked in the emerging field of big data.


## 1. Introduction

Big data is heralded as a powerful new resource for social science research. The excitement around big data emerges from the recognition of the opportunities it may offer to advance our understanding of human behaviour and social phenomenon in a way that has never been possible before (see for example Burrows and Savage, 2014; Kitchin, 2014a,b; Manovich, 2011; Schroeder, 2014). The concept of big data is vague however and has never been clearly defined (Harford, 2014a,b). We contend that this is highly problematic and leads to unnecessary confusion. Multiple definitions of big data are available and many of these seem to unwittingly focus on one specific type of data (e.g. social media data or business data) without appreciating the differences between the various types of data which could also reasonably be described as big data. We argue that there are multiple types of big data and that each of these offer new opportunities in specific areas of social investigation. These different types of big data will often require different analytical approaches and therefore a clearer understanding of the specific nature of the data is vital for undertaking appropriate analyses.

We highlight that whilst there may be a 'big data revolution' underway, it is not the size or quantity of these data that is revolutionary. The revolution centres on the increased availability of new types of data which have not previously been available for social science research. By treating big data as a single unified entity social scientists might fail to adequately

* Corresponding author.
*E-mail addresses:* R.Connelly@warwick.ac.uk (R. Connelly), chris.playford@ed.ac.uk (C.J. Playford), Vernon.Gayle@ed.ac.uk (V. Gayle), chris.dibben@ed.ac.uk (C. Dibben).

http://dx.doi.org/10.1016/j.ssresearch.2016.04.015

appreciate the attributes and potential research value of these new data resources. We argue that careful consideration of these different types of data is required to avert the risk that researchers will miss valuable data resources in the rush to exploit data with the highest profile.

In this paper we aim to provide a thorough treatment of administrative data which is one particular type of big data. Administrative data can be generally described as data which are derived from the operation of administrative systems (e.g. data collected by government agencies for the purposes of registration, transaction and record keeping) (Elias, 2014). We emphasise this form of big data for two reasons. First, we observe that administrative data has been largely neglected from many of the mainstream discussions of big data. Second, because administrative data are particularly valuable and may provide the means to address fundamental questions in the social sciences and contribute directly to the evidence base (e.g. answering questions relating to social inequality). This paper begins with a review of available definitions of big data, and we emphasise why administrative data should be characterised as a form of big data. We then consider how administrative data compares with the traditional types of data used in the social sciences (e.g. social survey data). Finally, we discuss the opportunities and challenges offered by the use of administrative data resources in social science research.

## 2. What is big data?

There is no single clear definition of big data. de Goes (2013) has gone as far as to suggest that the term big data is too vague and wide-ranging to be meaningful. In this section we summarise some of the definitions of big data in an attempt to bring more clarity to what big data constitutes.

Taylor et al. (2014) conducted a series of interviews with high profile economists working in this field in an attempt to better understand big data and its uses. These researchers identified the size and complexity of datasets as a key component of big data. Centrally, they emphasised that the increased number of observations and variables available in datasets were the result of a shift in the sources of data which were available to them (especially from the internet and social media). Generally the size and coverage of datasets are a central element of the definition of big data. Einav and Levin (2013) also emphasise that data is now available faster, and has a far greater coverage than the data resources which were previously available to social researchers.

Much of the literature discussing big data focuses on data which results from online activities and the use of social media (see for example Tinati et al., 2014). This type of data may be produced through online searches, internet viewing histories, blogs, social media such as Twitter and Facebook posts, and the sharing of videos and pictures.[1] The growth of the internet and electronic social networking has resulted in the unprecedented collection of vast amounts of data. The use of internet and social media data have resulted in numerous research studies investigating a wide range of topics such as individual's moods (e.g. Dodds et al., 2011), politician's impression management (e.g. Jackson and Lilleker, 2011) and collective political action (e.g. Segerberg and Bennett, 2011).

Big data should not be considered as synonymous with data collected through the internet. This is because big data can also originate from sources such as commercial transactions, for example purchases in-store from supermarkets or from bank transactions (see Felgate and Fearne, 2015). Big data can originate from sensors, for example satellite and GPS tracking data from mobile phones (see Eagle et al., 2009). Genome data is a source of big data and programs such as the '100,000 Genomes Project' in the UK and the 'Precision Medicine Initiative' in the US have resulted in the collection of massive amounts of data for the purpose of genome sequencing (see Eisenstein, 2015). Administrative data, for example education records, medical records, and tax records, are also sources of big data (see Chetty et al., 2011a,b).

Perhaps the most well-known definition of big data is provided by Laney (2001), who describes big data in terms of volume (i.e. the amount of data), variety (i.e. the range of data formats available such as text, pictures, video, financial or social transactions), and velocity (i.e. the speed of data generation). Tinati et al. (2014) highlight the all-encompassing nature of big data (i.e. the data captures all of the information from a particular platform such as twitter). Tinati et al. (2014) also consider the real-time nature of big data as one of its key features (i.e. big data may be captured on events or interactions as they happen).

Schroeder and Cowls (2014) emphasise that the concept of big data is strongly associated with a step-change in the types of data resources which are becoming available to researchers. Similarly, Harford (2014a,b) highlights that the 'found' nature of big data is one of its fundamental features. In the era of big data we are increasingly dealing with data resources that have been discovered by researchers as potential sources of valuable research data, but which have been collected for different (i.e. non-research) purposes. Traditional sources of data in the social sciences are 'made' by researchers. Even large scale social survey data resources which are used by many researchers, who are often working in different fields, to answer different questions are designed specifically for research purposes. By contrast big data are data resources which were collected for purposes other than research and researchers do not have any input into the design of these data or its content. A central characteristic that could be added to a definition of big data is that these data are not collected for research purposes but can be suitably re-purposed by social science researchers.

[1] This type of internet data is distinct from the exhaust data generated as trails of information created as a by-product resulting from internet or online activities (e.g. log files, cookies, temporary files). Exhaust data is of value to marketers and businesses (see Ohlhorst, 2012), however it tends to have less relevance for social science research.

Please cite this article in press as: Connelly, R., et al., The role of administrative data in the big data revolution in social science research, Social Science Research (2016), http://dx.doi.org/10.1016/j.ssresearch.2016.04.015

The term big data is often used to describe the innovative large-scale sources of quantitative data which are increasingly becoming available for research purposes, however this term does not describe a coherent and uniform set of data resources. Kitchin (2014a,b) highlights that there are many attributes that have been suggested to characterise big data, however more work needs to be done to establish different varieties of data under the big data umbrella. Kitchin (2014a,b) emphasises that some data may hold many of the characteristics thought to fulfil the definition of big data (e.g. size, variety and velocity) and other types of big data may hold a different set of characteristics, or only a single big data characteristic, but can still be considered as a form of big data. Not all big data resources will be equally large, not all will involve fast and real-time data availability, and not all will include an all-encompassing wide range of information. We suggest that a unifying characteristic of big data resources are that they are found data rather than made data. Big data also represents a development in the accessibility of certain, largely quantitative, data resources which have been more restricted to social science researchers in the past.

Kitchin (2014b) advocates the need to produce a taxonomy of big data with clear examples of particular data types. With this in mind in the following section we describe administrative data. We then outline how this specific form of data should be considered as one type of big data.

## 3. Administrative data

Administrative data are defined as data which derive from the operation of administrative systems, typically by public sector agencies (see Elias, 2014). Similarly, Woollard (2014) summarises administrative data as information collected for the purposes of registration, transaction and record keeping, and administrative data are often associated with the delivery of a service. These data can be derived from a wide range of administrative systems such as those used in education, healthcare, taxation, housing, or vehicle licensing. Administrative data also include information from registers such as notifications of births, deaths and marriages, electoral registration, and national censuses. Although administrative data have not been central to discussions of big data, we consider that these data fit firmly within the definitions of big data described above.

Administrative data are a source of large and complex quantitative information, and they are found data that are primarily generated for a purpose other than research. In some nations, such as Norway, Finland and Sweden, administrative data resources have been available to researchers for many years (see United Nations, 2007). In other parts of the world, especially the UK and the US, the recent increased availability of administrative data for research represents a step-change in the social science data infrastructure.

## 4. Administrative social science data and traditional sources of social science data

The main distinction between administrative social science data[2] and the data traditionally used in the social sciences, can be effectively characterised by the distinction between found and made data. Generally, social scientists make use of made data which they collect through experiments and observational studies (e.g. social surveys). The main distinctions between the characteristics of traditional social science data resources, administrative social science data, and other types of big data are outlined in Fig. 1.

### 4.1. Made data

Made data collected through experimental methods are designed and collected to address well defined hypotheses (see first panel of Fig. 1). These data may be large and complex, but they are usually smaller than administrative social science data sources. These data are highly systematic (i.e. structured and clearly organised). The researcher also has clear information on the sample that these data come from. We can therefore be clear about the representativeness of these data and the possibilities for wider inferences. As these data are collected specifically for the analysis of a small number of hypotheses, the re-use value of these data by other researchers may be relatively limited.

The second panel of Fig. 1 describes observational data. This is perhaps the most widely used type of data in social science research and includes the large multipurpose social survey datasets that are used extensively in quantitative social research (e.g. the European Social Attitudes Survey, The German Socio-Economic Panel, The Panel Study of Income Dynamics, and the UK Millennium Cohort Study). In contrast to experimental data these resources are not collected with the aim of answering a single clearly defined research question. These infrastructural data resources are designed by researchers to provide the information required to answer a wide range of research questions. Because of the general nature of these data resources they have high re-use potential in social science research.

As large multipurpose social survey datasets are designed for the purpose of research, a great deal of thought is placed into how information is collected to ensure that the data produced is of the highest quality, is suitable for research purposes, and that the measures included are valid and reliable. These data resources are usually large and complex, and the data collection

[2] We use the term 'administrative social science data' to describe administrative datasets that contain information relevant to social science research. We use this term to make a distinction between other types of administrative data that may have value for operational research, but are unlikely to be used for social and economic research.

Please cite this article in press as: Connelly, R., et al., The role of administrative data in the big data revolution in social science research, Social Science Research (2016), http://dx.doi.org/10.1016/j.ssresearch.2016.04.015

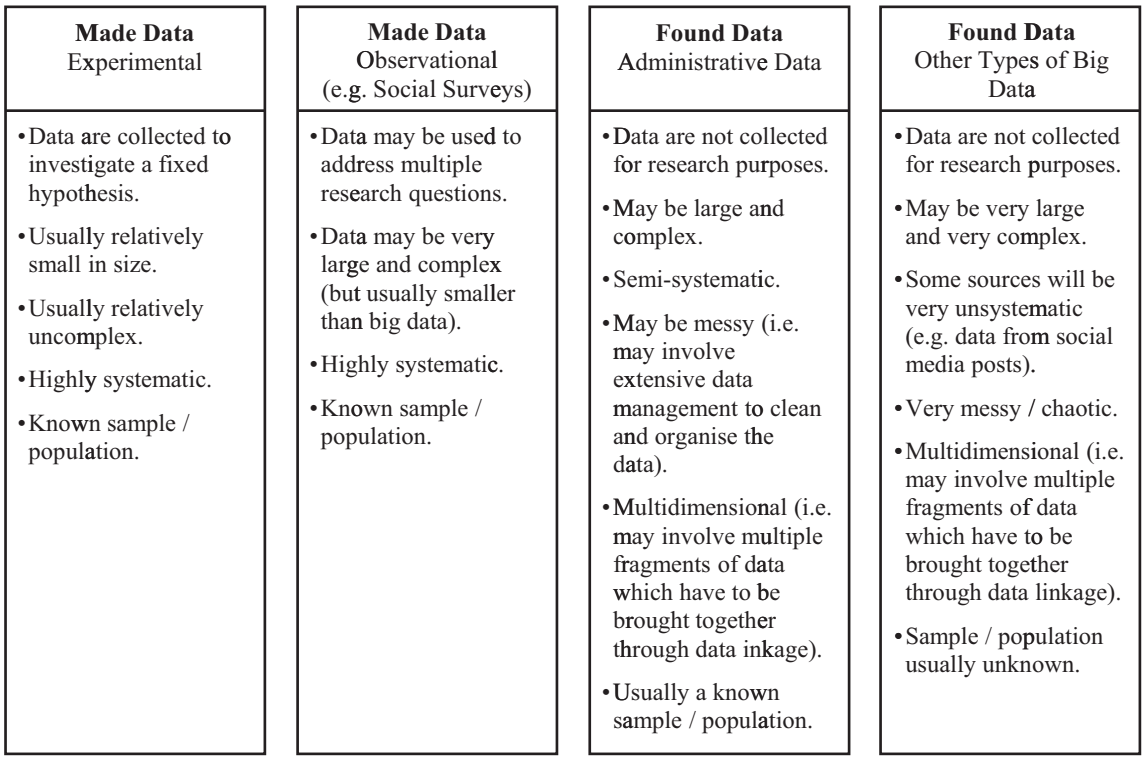


**Fig. 1.** Characteristics of quantitative social science data resources.

is highly systematic. Large scale social survey data collectors will typically employ a large team to clean, organise and document the data. These datasets are available in formats that facilitate the widest possible range of social science analyses. Like experimental data, observational studies collect information from a well-planned sample of individuals (or households) from known populations so that researchers can make general inferences about wider society.

### *4.2. Administrative social science data*

As we have described, administrative social science data differs from made data as it is not originally collected for the purpose of research (see panel 3 of Fig. 1). Researchers generally have no input into the design, structure and content of administrative social science data. These datasets can be large, however they are often not as large as the types of big data collected through for example, social media, GPS tracking, or supermarket transactions. These data are also likely to be more complex than the well curated social survey data resources which researchers may be accustomed to. The data may be messy and the use of administrative social science data resources is likely to involve substantial data management (or enabling) to clean and organise the data into the format required for analysis.

Many administrative social science datasets will be multidimensional in nature (i.e. the data will come in the form of fragments from different administrative systems). This multidimensional character is typically achieved through data linkage to join together datasets to gain all the pieces of information required to answer a social science research question (see Elias, 2014). Administrative data will generally be drawn from a known population, and will often retrieve information from an entire population rather than a sample (e.g. all 16 year olds completing school examinations in a given year). Often the administrative data available will represent specific populations, for example welfare benefits records will only include information on those individuals who claimed benefits, and this may restrict the focus of some research questions. When using administrative social science data researchers should, with some careful investigative work, have a clear idea of the population which the data covers.

In many respects the characteristics of administrative social science data make it relatively straightforward to incorporate into existing data analysis methodologies and epistemological data analysis perspectives. Concomitantly, the lessons learned from decades of data analysis and methodological work in areas such as social statistics, econometrics and sociology are still highly relevant for the analysis of administrative social science data.

Please cite this article in press as: Connelly, R., et al., The role of administrative data in the big data revolution in social science research, Social Science Research (2016), http://dx.doi.org/10.1016/j.ssresearch.2016.04.015

### 4.3. Other types of big data

Other forms of big data can be extremely large and complex, very unsystematic, messy and even chaotic (see panel 4 of Fig. 1). These data can also be drawn from unknown populations, and may involve complex or unknown samples. These characteristics have caused some social scientists using big data to re-evaluate their epistemological standpoints and also adopt new methodological techniques for analysing these data (see for example Kitchin, 2014a,b). It is probably too early to assess whether or not this re-evaluation is widespread or will have longer term consequences. Our position is that the samples and data structures of administrative social science data resources mean that these resources can be largely incorporated into the existing methodological traditions of social science data analysis (see Harford, 2014a,b). This is in contrast to other forms of big data.

In many ways one of the key benefits of administrative social science data is that it will be complementary to sources of made data (e.g. social survey data). Social survey data can provide the means to collect detailed information not available in administrative data. Administrative data can provide independent measures and additional information (e.g. educational examination results, medical conditions or tax records). Administrative data are especially powerful for collecting information that is more difficult to collect with a high degree of accuracy in a social survey context (e.g. the exact start and end dates of a job). In addition the linkage of these made and found data resources will greatly reduce the burden on survey respondents.

## 5. The opportunities arising out of administrative social science data

Administrative social science data are powerful resources, particularly because of the insights these data might offer into social inequality, human behaviours and the effectiveness of social policies (Card et al., 2010; Einav and Levin, 2013). Social researchers increasingly have access to large-scale high quality social survey data, however these data cannot include the information required to study all societal phenomenon. In some instances even social surveys that have large samples cannot support robust statistical analyses of specific sub-samples, even when the study has been designed with an element of over-sampling (which is sometimes referred to as 'boosting').

Administrative social science data will generally provide much larger sample sizes than social surveys, and at times administrative social science data may cover the entire population of interest (Card et al., 2010). This is sometimes referred to as $n =$ all. This has obvious advantages for the study of small subgroups and certain rare events.[3] Administrative data may also provide a means to access information on those groups who may be the least likely to take part in primary social science research (e.g. individuals from disadvantaged social groups). The large sample sizes of administrative social science datasets may enable the use of novel analytical approaches (e.g. quasi-causal methods) which take advantage of variations in the experiences of groups within these large samples (see Dunning, 2012). Administrative social science data may also be particularly useful for studying issues that individuals might be reticent to disclose to a primary researcher (e.g. mental health problems or substance abuse) (Goerge and Lee, 2001).

Administrative social science data offers the possibility for creating cohorts of individuals to study change over time and to pull together information on individuals who experienced a particular historical event (e.g. a major recession, or a change in educational examination systems) where there was no primary data collection at the time. In the collection of primary observational data social scientists have placed great value on cohort data, particularly birth cohort data (for example see Connelly and Platt, 2014; Elliott and Shepherd, 2006; Power and Elliott, 2006; Wadsworth et al., 2006). Cohort studies facilitate the examination of processes of societal change and theoretically allow researchers to distinguish between age and cohort effects (see Dale and Davies, 1994). The construction of cohorts using administrative data helps to facilitate the study of longitudinal processes and social change over time. Picot and Piraino (2012) demonstrate the value of producing cohorts using administrative data in their analysis of immigrant earnings in Canada. They utilise longitudinal tax data linked to immigrant landing records in order to effectively estimate the change in immigrant earnings and the immigrant-Canadian-born earnings gap for a series of cohorts. This is an example of where administrative records may facilitate research where there is an absence of social survey data.

As well as providing new research opportunities and filling gaps in the availability in primary research data, administrative social science data may also offer savings in comparison to the costs of primary data collection (see Zhang, 2012). At the current time, particularly in the UK, government researchers do not use their own data to their full potential. The increased use of administrative social science data, improved access to administrative social science data for academic researchers, and improved facilities for data linkage all have the potential for long-term data production cost efficiencies. Increased use of administrative social science data could also reduce the burdens on those individuals who take-part in primary data collections.

Administrative social science data may be particularly valuable for the evaluation of social policies and policy relevant issues, and the analysis of administrative social science data may contribute to the development of social policies. Examples of

[3] It is possible that administrative data may enable the study of small groups. However in small countries, with relatively small population sizes (e.g. Scotland), researchers should remain cognisant of the fact that sample sizes for very specific subgroups may still be too small for robust multivariate analyses.

Please cite this article in press as: Connelly, R., et al., The role of administrative data in the big data revolution in social science research, Social Science Research (2016), http://dx.doi.org/10.1016/j.ssresearch.2016.04.015

the use of administrative data for policy relevant research can be found in a number of substantive areas for example neighbourhood characteristics and safety (see O'Brien et al., 2015), children at risk of poor outcomes as young adults (see Crichton et al., 2015), the earnings of university graduates (see Britton et al., 2015), and the association between the employment of ex-offenders and their recidivism (see Justice, 2011). Some of the most impressive examples of policy relevant and impactful research using administrative social science data come from the work of Raj Chetty and colleagues.

Chetty, et al. (2011a,b) investigated the long term impact of an educational experiment using administrative records. In seventy nine Tennessee schools from 1985 to 1989, a large sample of children were randomly assigned to classrooms with different characteristics from kindergarten to the third grade. Some children were assigned to small classes, some to large classes, and the characteristics of teachers (e.g. years of experience) were recorded. Chetty, et al. (2011a,b) linked this experimental data to administrative records to examine the effects which these educational experiences had on outcomes in adulthood. They found that kindergarten test scores were highly correlated with outcomes such as earnings at age 27, college attendance, home ownership and retirement savings. They were also able to identify that pupils in smaller kindergarten classes were more likely to attend college, and that pupils who had a more experienced teacher in kindergarten had higher earnings at age 27.

In more recent work, Chetty et al. (2014) examined intergenerational earnings mobility in the U.S. using administrative tax records. They constructed a series of cohorts with earnings data on tax records for both parents and their children in adulthood. They found that mobility remained extremely stable for cohorts of people born from 1971 to 1993. Overall Chetty et al. (2014) conclude that young people entering the labour market today have the same chances of moving up the income distribution (relative to their parents) as children born in the 1970s. These studies are international examples of the possibilities that administrative social science data open up for research in a broad range of areas. In the next section we turn our attention to the challenges that surround the use of administrative social science data in research.

## 6. The challenges of administrative social science data

The most pertinent issues concerning the use of administrative social science data are legal and ethical. As administrative data are not primarily collected for research purposes, the public may have concerns over their privacy, the linkage of their data from different sources, and the use of their data by researchers. We do not consider these issues in this paper, but they have been discussed in depth elsewhere (see for example Stevens and Laurie, 2014). It is important to note that researchers who use administrative social science data will be working within a strict set of conditions given by the data owners (e.g. government departments). These conditions currently include undertaking specialised training, accessing the data from a secure setting where the data use is controlled and monitored, and having research outputs checked to ensure that individuals cannot be identified and information on individuals cannot be disclosed. Administrative social science datasets will be constructed so that individuals or households cannot be identified. This ensures that individuals' privacy will not be infringed by social science researchers using administrative data.

### 6.1. Data analysis

When administrative social science data are organised for conventional social science research (e.g. the application of multivariate techniques such as statistical models) they are indistinguishable from the familiar rectangular variable by case matrices from conventional social surveys, where a variable is recorded in each column and each case is allocated to a row.[4] Ultimately the challenges we will face in the analysis of administrative social science data are indistinguishable from those faced in the analysis of large-scale social surveys. For example, if an administrative dataset has repeated measurements on the same individuals the data will not be independent and identically distributed (Baltagi, 2008). To tackle data analysis challenges administrative data analysts should not ignore the helpful lessons that have emerged from decades of statistics and econometrics.

In most instances social science administrative datasets are observational, and they are not collected as part of an overall experimental design or experimental data collection protocol. The challenges associated with drawing causal inferences from observational social science data are well known (see Manski, 1993; Winship and Morgan, 1999). Multivariate methods, for example from the generalized linear mixed model family, provide improved statistical control for analyses of observational social science data (see Hedeker, 2005). We conjecture that statistical models (e.g. regression models) of observational data are best considered as 'sophisticated descriptions'.[5] There are a set of techniques emerging from econometrics which are aimed at providing partial solutions to the challenges of drawing inference from observational social science data (for an accessible review see Angrist and Pischke, 2008).

We can conceive of some circumstances where the nature of the social science administrative data facilitate quasi-experimental analyses. For example geographic variations in policies or temporal differences between policy interventions, might possibly provide a natural experiment. In other situations matching techniques or discontinuity

[4] The variable by case matrix will be familiar to researchers who have been trained to undertake statistical analyses of social science data, and it is described in standard elementary textbooks, for example see De Vaus (2014).

[5] We are grateful to Professor Sascha O. Becker, University of Warwick for introducing us to this terminology.

Please cite this article in press as: Connelly, R., et al., The role of administrative data in the big data revolution in social science research, Social Science Research (2016), http://dx.doi.org/10.1016/j.ssresearch.2016.04.015

designs might also be possible using administrative social science data. The frequent lack of a suitable set of background or auxiliary measures within administrative datasets is likely to constrain the use of many of the more advanced statistical approaches that have currently been proposed to assist researchers in drawing causal inference from observational data.

The impact of the size of administrative datasets is a particular issue that researchers should be cognisant of. Nevertheless administrative data are characterised as having a large $n$, but a relatively small number of variables ($k$). Therefore the dataset's dimension is less of a concern in administrative data analysis, compared with the analysis of other forms of 'big' data that have both large $n$ and $k$. Specialised statistical data analysis packages such as Stata and R are more than capable of handling datasets with a large number of observations. Stata MP, run on a modern multi-processor computer, can analyse 10 to 20 billion observations[6] with thousands of variables. This is more than adequate to analyse a dataset containing an observation for each of the world's population. On a practical level, the analysis of very large datasets will undoubtedly require powerful computing capacity, and the time taken for some computations to be completed may be extended.

If facing computing problems in the analysis of large datasets, a naive solution is to conduct the analysis on smaller samples of the data. Researchers must consider the effects on results of working with a reduced dataset. Researchers should also be suitably cautions to ensure that sub-samples appropriately represent underlying structures in the main dataset. When working with subsets of data we would recommend that researchers undertake sensitivity analyses and make these results transparent. It may also prove useful for researchers to consider insights from the methodological literature on sampling and re-sampling when undertaking analyses with subsets of large datasets (e.g. Kish, 1965). Another alternative is that researchers may resort to using high performance computing (e.g. supercomputers) that are equipped with very high speed processors and large memory capacities.

Importantly, from a statistical perspective the analysis of large datasets increases the need for researchers to be aware of the limitations of simplistic significance tests, and their associated $p$ values. For example when a standard linear regression model is estimated using an extremely large dataset the standard error calculated for a beta can be very small. Using a conventional statistical test of significance this can result in the potentially misleading conclusion that the effect of the variable (net of the other variables in the model) is very important[7] (see Lin et al., 2013; Sullivan and Feinn, 2012). This is not so much a problem of large samples, but a problem with how significance tests have come to be regarded in social science research. Discontent over the use of significance tests, is growing and is not specific to the analysis of large samples (see Carver, 1978; Gorard, 2015a,b; Johnson, 1999; Kühberger et al., 2014). A possible solution is to reduce the significance threshold of a $p$ value as the sample size increases (see Greene, 2003; Leamer, 1978). This is not a common practice and there are no accepted guidelines for acceptable significance levels at different sample sizes.

There are relatively straightforward steps that researchers can take to move beyond simply interpreting a $p$ value, for example researchers can present a measure of effect size alongside a $p$ value, such as a marginal effect (see Connelly et al., Forthcoming, Long and Freese, 2014; Vittinghoff et al., 2005). This would allow the researcher to demonstrate whether the finding has both statistical significance and substantive importance (i.e. how large a difference the observed effect might have on the outcome of interest). Demonstrating the substantive importance of results may be a particularly critical practice in administrative social science data analysis, due to the policy relevance of many of the issues that will be investigated using these data.

### 6.2. Data management

We have noted that, in contrast to traditional types of social science data, administrative social science data may be less systematic and require more data enabling by researchers to facilitate data analysis (see Einav and Levin, 2013; Goerge and Lee, 2001). Data enabling comprises tasks associated with preparing and shaping data for analysis. These tasks include re-structuring datasets, and recoding and constructing variables (see Long, 2009; Mitchell, 2010). When using administrative social science data, the process of data management may be more complicated and time demanding than when using a social survey that has been primarily collected for social science analyses and has been enabled by data collectors or by a social science data archive or provider. The time demands associated with enabling administrative data for social research will be especially burdensome when researchers are using large volumes of data from multiple administrative systems. Researchers will generally need to restructure their data to achieve the standard variable by case matrix that is required by most analysis techniques (Einav and Levin, 2013; Goerge and Lee, 2001). These data enabling skills are not generally taught to social scientists in any depth at the current time. Less experienced social science data analysts may struggle with this aspect of administrative data analysis, more than they would when using well curated social survey data resources.

Due to the lack of clear documentation **accompanying** many administrative social science data resources, researchers will need to exert time and effort to understand what types of questions could feasibly be answered using the administrative data

[6] Stata notes that the Stata MP software is ready to analyse up to 281 trillion observations once computer hardware catches up: http://www.stata.com/products/which-stata-is-right-for-me/.

[7] The p value problem will also affect tests of model assumptions, such as the Breusch-Pagan test of heteroscedasticity and the Durbin-Watson test of serial correlation. In very large samples these tests will tend to indicate a violation of regression assumptions, even for very small deviations so should be interpreted with suitable caution (see Lin et al., 2013).

Please cite this article in press as: Connelly, R., et al., The role of administrative data in the big data revolution in social science research, Social Science Research (2016), http://dx.doi.org/10.1016/j.ssresearch.2016.04.015

resources available. When using social survey data, researchers will often have access to a wealth of data documentation. Researchers will also be able to examine these data and conduct exploratory analyses to investigate the feasibility of their research before they commence with a full project. In the case of administrative social science data, where data access is restricted, the usual and necessary exploratory stage of the research process may not be easily practicable (Einav and Levin, 2013). Researchers may have to begin a research project without full knowledge of the data which they are using. Whilst this work may be hard for the pioneers, this challenge may be overcome as researchers share and document knowledge of the characteristics, advantages and drawbacks of particular administrative data resources.

### 6.3. Data generation process

Administrative social science data were not collected for research purposes and therefore are not documented in order to support data analysis. In order to fully understand administrative data records, researchers must put effort into uncovering the data generation processes which have governed how these data have been created. Gomm (2008) highlights that ethno-statistics may play an increasingly important role in the effective use of administrative social science data. Ethno-statistics are studies which have traditionally been conducted to understand the way facts about people are socially produced (Black, 1994; De Zwart, 2012; Gephart, 2006). Ethno-statistics studies have the potential to be extremely valuable in investigating and documenting the process through which administrative social science data are produced. These studies could assist researchers in evaluating the accuracy and consistency of measures, and provide better understanding of the intricacies of the data. Ethno-statistics studies have the potential to provide critical insights into possible errors and biases in administrative social science data resources.

Elliott (2015) has recounted the value of this type of investigative work when undertaking administrative data analysis of unemployment rates in Cambridge, England. She observed strangely high rates of unemployment in a particularly affluent area. Through discussing this issue with data collectors she was able to determine that the home address of homeless individuals had been recorded at the Job Centre (i.e. government employment agency) located in this area resulting in misleading unemployment figures. This is a clear example of where the investigation of data provenance proves to be critical. Without investigation into the data collection process the researchers would not have appreciated this data recording decision, and would ultimately have reported erroneous research findings.

Gomm (2008) also describes the importance of understanding the units included in an administrative dataset. He gives the example of counting the number of psychiatric in-patients over a period of a year. This may at first seem like a straightforward measure of the extent of severe mental illness. In-patient psychiatric care comes at the end of a process involving an individual's decision to consult a doctor, a doctor's diagnostic behaviour, and the diagnostic behaviour of the psychiatrist which an individual may or may not have been referred to. Therefore there are multiple interactions within the process of an individual reaching in-patient psychiatric care which mean that the simple count of psychiatric in-patients may not be either a reliable or a valid measure of severe mental illness.

Administrative social science data will also reflect a system that is itself responding to an ever changing policy context. Changes in administrative systems (e.g. changes in the financial assistance available to unemployed individuals, or the examinations undertaken in schools) may lead to changes in measures. Therefore it is highly desirable that researchers develop a clear biographical understanding of the administrative system in order to appreciate how information is collected and how measures are developed, at the same time as understanding how cases are included in datasets and how data may change over time. This may be a complicated process and will require researchers to build a detailed knowledge of both operational changes in data collection and policy changes in their field of analysis.

Marsh and Elliott (2008) greatly emphasise the detective work which social researchers must undertake when sifting through and piecing together numerical evidence about the social world (see also Tukey, 1977). The detective work of social science data analysis has never been more important than when analysing administrative social science data, which has not been collected and curated for the purposes of research. Understanding the processes and procedures through which administrative social science data have been created is central to understanding what these data represent and which units are included. In the spirit of cumulative social science endeavour, and to ensure that data reach their full analytical potential it is also important that researchers document their findings and make the processes by which administrative social science data have been generated transparent. This will directly benefit future researchers.

### 6.4. Data quality

Researchers should question the quality of the content of administrative social science data resources. In social survey research, analysts place careful attention on the measurement of their variables and studies continually question the accuracy, validity and reliability of the information collected (see for example Burton et al., 2010; Moore et al., 2000; Webb et al., 1999). The total error paradigm (see Groves et al., 2011) identifies multiple sources of error in social surveys including, measurement error, processing error, coverage error, sampling error, nonresponse error, and adjustment error. Administrative data may also have many of these errors, particularly measurement error, processing error, nonresponse error and adjustment error, although administrative data are generally less likely to contain coverage error and, because they are not based on samples, sampling error (Groen, 2012).

Please cite this article in press as: Connelly, R., et al., The role of administrative data in the big data revolution in social science research, Social Science Research (2016), http://dx.doi.org/10.1016/j.ssresearch.2016.04.015

Goerge and Lee (2001) emphasise that the degree of error varies between administrative data systems, and between items within an administrative data system. Therefore they encourage researchers to assess each new administrative social science dataset individually for each new research question at hand. Goerge and Lee (2001) encourage researchers to question the original motivation for collecting the administrative data. This can prove to be consequential, for example data collected for financial purposes often produces the most reliable information as there were clear incentives for the data collectors to ensure that the data recorded was accurate. However, the inaccuracies that individuals detect and the errors that emerge in transactions with welfare benefits agencies, tax authorities, transport agencies, health services and local authorities, coupled with the errors, miscalculations and inaccuracies that occur in transactions with service providers such as banks, credit card companies, utility companies, and delivery and transport providers, should all hang a reasonable question mark over the quality of some administrative data for social research.

Importantly researchers should also consider whether the information they are interested in is central to the purposes it was collected for. If certain measures are not required for the operation of an administrative system they may not be collected diligently (Goerge and Lee, 2001). For example, Goerge et al. (1992) studied foster care workers who were asked to provide disability details of children on a computerised record. In the vast majority of cases this information had no impact on the actions or decisions of the workers, therefore the data collected within the administrative system was of poor quality. For an excellent account of the influence that workers may have within administrative systems see Lipsky (1979).

Goerge and Lee (2001) also provide a summary of some of the practical strategies that should be used to assess the quality of administrative social science data resources. First, the most prosaic of these is to compare the administrative data with another source, although in practice this is not always possible. Second, researchers should enquire as to whether there was a system of auditing the data (e.g. has the data been cross-checked at any point?). Third, was the data entered by a frontline worker? Passing on data to a data entry clerk provides an additional opportunity to introduce errors and prevents the frontline worker from seeing incorrect data and correcting it. Fourth, did quality assurance checks exist within the data collection system? For example was the data entry system programmed to reject invalid values or empty fields?

A further important consideration in assessing data quality is whether the administrative data is associated with some performance management system or target. According to Campbell's Law[8] when a measure is used for social decision making it will become increasingly susceptible to corruption (Campbell, 1979). Therefore the value of the measure for monitoring the social phenomenon under investigation will be significantly reduced. One recent high profile example of this principle relates to standardised educational testing in the US (Nichols and Berliner, 2005). Standardised test scores in the US are used to judge the performance of schools, they can influence the employability of individual teachers and administrators, and they can influence teachers' bonus pay. These tests can also determine the promotion or non-promotion of a pupil to a higher school grade, or the attainment of a high school diploma. The high-stakes attached to these tests have led to widely reported changes in the actions of individuals associated with these tests. These include examples of administrators and teachers cheating, pupils cheating, the exclusion of low-performing pupils from the opportunity to take the test, and teachers specifically training pupils to pass these tests (see Nichols and Berliner, 2005). Researchers should therefore question whether the high-stakes associated with these educational tests have influenced their validity as a measure of school pupils' educational performance.

### 6.5. Data access

A further challenge, which has been eluded to above, results from the fact that these data do not belong to the research community and therefore gaining access to these data can be very tricky and time consuming. Access issues limit the researcher in planning their research, often prevent exploratory data analysis, and limit the extent to which new cohorts of social scientists can be trained in the use of these data (Einav and Levin, 2013). The use of secure data, which cannot necessarily be accessed by the wider research community also has implications for replication and the development of cumulative social science. We contend that one of the ways in which these issues could be ameliorated is with the production of clear and assiduous documentation of the research process (see Dale, 2006; Freese, 2007; Long, 2009; Steuer and Binmore, 2003). In particular we advocate making data analysis syntax files accessible.[9] The use of this documentation and accessible syntax files would allow other researchers to scrutinise analyses and build upon existing work in the future.

### 6.6. Data linkage

In some northern European countries, such as Norway, Finland and Sweden an effective infrastructure is already in place for the linkage and sharing of administrative data for research purposes (see United Nations, 2007). The availability of

[8] Also known as Goodhart's Law (see Chrystal and Mizen, 2003; Elton, 2004; Goodhart, 1984).

[9] The term 'syntax file' originates from SPSS, however we use the term to refer to all command files in statistical data analysis software packages such as. do files in Stata. We also consider R scripts to be syntax files because similarly they command data analyses.

Please cite this article in press as: Connelly, R., et al., The role of administrative data in the big data revolution in social science research, Social Science Research (2016), http://dx.doi.org/10.1016/j.ssresearch.2016.04.015

consistent unique identification numbers in these countries is a major advantage for the facilitation of administrative data research (Einav and Levin, 2013). In countries like the UK and the US the lack of consistent identification numbers across administrative systems means that linking together the pieces of information required to answer a social research question is usually much more complex.

With consistent unique identifiers deterministic data linkage is possible, this is because there is a perfect match between individuals on two datasets. Without unique identifiers researchers may have to rely on other methods such as probabilistic linkage based on a series of weaker identifiers (e.g. name, date of birth, gender). With probabilistic data linkage there is a risk that linkage is not accurate, introducing error into the analysis (Goerge and Lee, 2001). There have been statistical developments made in dealing with possible linkage errors when using linked data (see Belin and Rubin, 1995; Lahiri and Larsen, 2005; Scheuren and Winkler, 1993). The accuracy of probabilistic data linkage will vary from dataset to dataset, and some administrative datasets will completely lack the identifiers required to enable linkage. Administrative data researchers must carefully consider the quality of the linkage between their datasets, and the influence which this may have on their substantive results.

## 7. Conclusions

Big data has the potential to change the landscape of social science research, and administrative social science data may offer particular benefits. We have highlighted that although big data is currently a popular buzzword in social science, big data is not a unified type of data resource but a multifaceted collection of data types. The lack of a clear definition of what big data is, presents a major impediment to the social science community whilst we are attempting to come to grips with the wide range of data resources which are increasingly becoming available. We consider that one of the central and defining features of big data is its found nature (i.e. big data are not collected for research purposes). Social science researchers must learn how to understand and handle these data resources to ensure that they reach their fullest analytical potential.

With the notable exceptions of the Nordic countries, access to administrative social science data is still highly restricted in many nations. This is perhaps the major impediment to realising the potential of administrative data research. Looking to the Nordic nations we can see that the use of consistent identifiers and effective data sharing infrastructures have facilitated the use of administrative social science data by researchers. Access to administrative social science data is more restricted in the USA (Card et al., 2010), Canada (Doiron et al., 2013), Western Australia (Holman et al., 2008) and the United Kingdom (Administrative Data Taskforce, 2012). As part of the big data revolution many countries are currently taking steps to improve access to administrative data for research purposes[10].

Administrative social science data offer the opportunity to study policy changes, social problems and societal issues using information which may not routinely be available in social surveys. The large size of many administrative social science data resources may offer the opportunity to study sub-groups, and could potentially lead to analytical approaches such as quasi-experimental methods being used more routinely. The re-purposing of these data could also result in long term savings for government departments, and social science data producers.

As administrative social science data are not collected for the purposes of research, these data are generally more messy and complex than traditional social science datasets. Researchers should therefore not underestimate the amount of time and effort they will need to spend undertaking data enabling tasks to prepare administrative data for analysis. The importance of developing good data enabling skills is often overlooked, especially when teaching data analysis to students. We suggest that these skills are important if social scientists are to effectively utilise administrative social science data.

The role of social scientists as data detectives is particularly important when analysing administrative data. Researchers will need to exert effort to build an understanding of the 'biography' of the administrative data they are using. Understanding the processes of how and why administrative data are collected will be central to assessing the data's quality and its suitability for social research. Big data has the potential to change the landscape of social science research, and administrative social science data offers particular benefits, however we conclude that for administrative social science data to have a suitably full impact there must be a step-change in research practices and research must routinely ensure that work undertaken using administrative social science datasets is efficient, transparent and reproducible.

## Acknowledgement

This work was supported by the Economic and Social Research Council [Grant Number ES/L007487/1].

## References

Administrative Data Taskforce, 2012. The UK Administrative Data Research Network: Improving Access for Research and Policy. Economic and Social Research Council, London.

[10] For example, in the UK, the Administrative Data Research Network (ADRN) has recently been set up to provide an Administrative Data Research Centre in each of the UK territories (http://adrn.ac.uk/). The aim of this network is to develop an infrastructure that assists researchers with gaining access to administrative data resources, provides services for linking different data sources, provides accreditation training for researchers, and provides secure facilities where data can be analysed.

Angrist, J.D., Pischke, J.-S., 2008. Mostly Harmless Econometrics: an Empiricist's Companion. Princeton university press.
Baltagi, B., 2008. Econometric Analysis of Panel Data. John Wiley & Sons, New York.
Belin, T.R., Rubin, D.B., 1995. A method for calibrating false-match rates in record linkage. J. Am. Stat. Assoc. 90, 694–707.
Black, N., 1994. Why we need qualitative research. J. Epidemiol. Community Health 48, 425.
Britton, J., Shephard, N., Vignoles, A., 2015. Comparing Sample Survey Measures of English Earnings of Graduates with Administrative Data during the Great Recession. Institute for Fiscal Studies, London.
Burrows, R., Savage, M., 2014. After the crisis? Big Data and the methodological challenges of empirical sociology. Big Data Soc. 1, 2053951714540280.
Burton, J., Nandi, A., Platt, L., 2010. Measuring ethnicity: challenges and opportunities for survey research. Ethn. Racial Stud. 33, 1332–1349.
Campbell, D.T., 1979. Assessing the impact of planned social change. Eval. Progr. Plan. 2, 67–90.
Card, D., Chetty, R., Feldstein, M.S., Saez, E., 2010. Expanding Access to Administrative Data for Research in the United States. American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas.
Carver, R., 1978. The case against statistical significance testing. Harv. Educ. Rev. 48, 378–399.
Chetty, R., Friedman, J.N., Hilger, N., Saez, E., Schanzenbach, D.W., Yagan, D., 2011a. How does your kindergarten classroom affect your earnings? evidence from project star. Q. J. Econ. 126, 1593–1660.
Chetty, R., Friedman, J.N., Rockoff, J.E., 2011b. The Long-term Impacts of Teachers: Teacher Value-added and Student Outcomes in Adulthood. National Bureau of Economic Research.
Chetty, R., Hendren, N., Kline, P., Saez, E., Turner, N., 2014. Is the United States still a land of opportunity? Recent trends in intergenerational mobility. Am. Econ. Rev. 104, 141–147.
Chrystal, K.A., Mizen, P.D., 2003. Goodhart's law: its origins, meaning and implications for monetary policy. In: Central banking, monetary theory and practice: Essays in honour of Charles Goodhart, 1, pp. 221–243.
Connelly, R., Platt, L., 2014. Cohort profile: UK millennium Cohort study (MCS). Int. J. Epidemiol. 43, 1719–1725.
Connelly, R., Gayle, V., Lambert, P., Statistical Modelling of Key Variables in Social Survey Data Analysis. Methodological Innovations. Forthcoming.
Crichton, S., Templeton, R., Tumen, S., 2015. Using Integrated Administrative Data to Understand Children at Risk of Poor Outcomes as Young Adults. New Zealand Treasury, Wellington, NZ.
Dale, A., 2006. Quality issues with survey research. Int. J. Soc. Res. Methodol. 9, 143–158.
Dale, A., Davies, R.B., 1994. Analyzing Social and Political Change: a Casebook of Methods. Sage.
de Goes, J., 2013. “Big Data” Is Dead. What's Next?, VB/Big Data.
De Vaus, D.A., 2014. Surveys in Social Research. Routledge, Abingdon, Oxon.
De Zwart, F., 2012. Pitfalls of top-down identity designation: ethno-statistics in the Netherlands. Comp. Eur. Polit. 10, 301–318.
Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., Danforth, C.M., 2011. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PloS one 6, e26752.
Doiron, D., Raina, P., Fortier, I., 2013. Linking Canadian population health data: maximizing the potential of cohort and administrative data. Can. J. Public Health 104, e258–e261.
Dunning, T., 2012. Natural Experiments in the Social Sciences: a Design-based Approach. Cambridge University Press.
Eagle, N., Pentland, A.S., Lazer, D., 2009. Inferring friendship network structure by using mobile phone data. Proc. Natl. Acad. Sci. 106, 15274–15278.
Einav, L., Levin, J.D., 2013. The Data Revolution and Economic Analysis. National Bureau of Economic Research.
Eisenstein, M., 2015. Big data: the power of petabytes. Nature 527, S2–S4.
Elias, P., 2014. Administrative data. In: Duşa, A., Nelle, D., Stock, G., Wagner, G. (Eds.), Facing the Future: European Research Infrastructures for the Humanities and Social Sciences. SCIVERO, Berlin, pp. 47–48.
Elliott, J., 2015. Advancing the Administrative Data Research Network: Next Steps for Facilitating Excellent Research, Administrative Data Research Network Annual Research Conference. Queen's University, Belfast.
Elliott, J., Shepherd, P., 2006. Cohort profile: 1970 British birth cohort (BCS70). Int. J. Epidemiol. 35, 836–843.
Elton, L., 2004. Goodhart's law and performance indicators in higher education. Eval. Res. Educ. 18, 120–128.
Felgate, M., Fearne, A., 2015. Analyzing the Impact of Supermarket Promotions: a Case Study Using Tesco Clubcard Data in the UK, the Sustainable Global Marketplace. Springer, pp. 471–475.
Freese, J., 2007. Replication standards for quantitative social science why not sociology? Sociol. Methods Res. 36, 153–172.
Gephart, R.P., 2006. Ethnostatistics and organizational research methodologies an introduction. Organ. Res. Methods 9, 417–431.
Goerge, R.M., Lee, B.J., 2001. Matching and cleaning administrative data. In: Citro, C.F., Moffitt, R.A., Van Ploeg, M. (Eds.), Studies of Higher Population: Data Collection and Research Issues. National Academies Press, Washington D.C., pp. 197–219
Goerge, R.M., Van Voorhis, J., Grant, S., Casey, K., 1992. Special-education experiences of foster children: an empirical study. Child Welf. J. Policy Pract. Progr. 71 (5), 419–437.
Gomm, R., 2008. Social Research Methodology: a Critical Introduction. Palgrave Macmillan.
Goodhart, C.A.E., 1984. Monetary Theory and Practice: the UK Experiencie. Macmillan Publishers Limited.
Gorard, S., 2015a. Rethinking ‘quantitative’methods and the development of new researchers. Rev. Educ. 3 (1), 72–96.
Gorard, S., 2015b. What to do instead of significance testing? calculating the'number of counterfactual cases needed to disturb a finding'. Int. J. Soc. Res. Methodol. 1–10. http://dx.doi.org/10.1080/13645579.2015.1091235.
Greene, W.H., 2003. Econometric Analysis. Pearson, Harlow.
Groen, J.A., 2012. Sources of error in survey and administrative data: the importance of reporting procedures. J. Off. Stat. 28, 173.
Groves, R.M., Fowler Jr., F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R., 2011. Survey Methodology. John Wiley & Sons.
Harford, T., 2014a. Big data: a big mistake? Significance 11, 14–19.
Harford, T., 2014b. Royal Statistical Society Significance Lecture. The Big Data Trap., Significance. Royal Statistical Society, London.
Hedeker, D., 2005. Generalized linear mixed models. In: Everitt, B., Howell, D. (Eds.), Encyclopaedia of Statistics in Behavioral Science. Wiley, New York, pp. 729–738.
Holman, C.D., Bass, A.J., Rosman, D.L., Smith, M.B., Semmens, J.B., Glasson, E.J., Brook, E.L., Trutwein, B., Rouse, I.L., Watson, C.R., de Klerk, N.H., Stanley, F.J., 2008. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. Aust. Health Rev. 32, 766–777.
Jackson, N., Lilleker, D., 2011. Microblogging, constituency service and impression management: UK MPs and the use of Twitter. J. Legis. Stud. 17, 86–105.
Johnson, D.H., 1999. The insignificance of statistical significance testing. J. Wildl. Manag. 763–772.
Justice, M.o., 2011. Offending, Employment and Benefits: Emerging Findings from the Data Linkage Project. Ministry of Justice, London.
Kish, L., 1965. Survey Sampling. Wiley, New York.
Kitchin, R., 2014a. Big Data, new epistemologies and paradigm shifts. Big Data Soc. 1, 2053951714528481.
Kitchin, R., 2014b. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. Sage, London.
Kühberger, A., Fritz, A., Scherndl, T., 2014. Publication Bias in Psychology: a Diagnosis Based on the Correlation between Effect Size and Sample Size.
Lahiri, P., Larsen, M.D., 2005. Regression analysis with linked data. J. Am. Stat. Assoc. 100, 222–230.
Laney, D., 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note 6.
Leamer, E.E., 1978. Specification Searches: Ad Hoc Inference with Nonexperimental Data. John Wiley & Sons, London.
Lin, M., Lucas, H.C., Shmueli, G., 2013. Research commentary-too big to fail: large samples and the p-value problem. Inf. Syst. Res. 24, 906–917.
Lipsky, M., 1979. Street Level Bureaucracy. Russell Sage Foundation, New York.
Long, J.S., 2009. The workflow of data analysis using Stata. Stata Press books.

Please cite this article in press as: Connelly, R., et al., The role of administrative data in the big data revolution in social science research, Social Science Research (2016), http://dx.doi.org/10.1016/j.ssresearch.2016.04.015

Long, J.S., Freese, J., 2014. Regression Models for Categorical Dependent Variables Using Stata. Stata Press, College Station.
Manovich, L., 2011. Trending: the Promises and the Challenges of Big Social Data.
Manski, C.F., 1993. Identification of endogenous social effects: the reflection problem. Rev. Econ. Stud. 60, 531–542.
Marsh, C., Elliott, J., 2008. Exploring Data. An Introduction to Data Analysis for Social Scientists Polity Press, Cambridge.
Mitchell, M.N., 2010. Data Management Using Stata: a Practical Handbook. Stata Press books.
Moore, J.C., Stinson, L.L., Welniak, E.J., 2000. Income measurement error in surveys: a review. J. Off. Stat. 16, 331–362.
Nichols, S.L., Berliner, D.C., 2005. The Inevitable Corruption of Indicators and Educators through High-stakes Testing. Education Policy Research Unit, Arizona State University.
O'Brien, D., Sampson, R., Winship, C., 2015. Ecometrics in the age of big data. Measuring and assessing “broken windows” using large-scale administrative records. Sociol. Methodol. 45, 101–147.
Ohlhorst, F.J., 2012. Big Data Analytics: Turning Big Data into Big Money. John Wiley & Sons.
Picot, G., Piraino, P., 2012. Immigrant Earnings Growth: Selection Bias or Real Progress? Statistics Canada Analytical Studies Branch Research Paper Series.
Power, C., Elliott, J., 2006. Cohort profile: 1958 british birth cohort (national child development study). Int. J. Epidemiol. 35, 34–41.
Scheuren, F., Winkler, W.E., 1993. Regression analysis of data files that are computer matched. Surv. Methodol. 19, 39–58.
Schroeder, R., 2014. Big Data and the brave new world of social media research. Big Data Soc. 1, 2053951714563194.
Schroeder, R., Cowls, J., 2014. Big Data, Ethics, and the Social Implications of Knowledge Production.
Segerberg, A., Bennett, W.L., 2011. Social media and the organization of collective action: using Twitter to explore the ecologies of two climate change protests. Commun. Rev. 14, 197–215.
Steuer, M., Binmore, K., 2003. The scientific study of society. Kluwer Academic Dordrecht.
Stevens, L.A., Laurie, G., 2014. The Administrative Data Research Centre Scotland: a Scoping Report on the Legal & Ethical Issues Arising from Access & Linkage of Administrative Data. Edinburgh School of Law Research Paper.
Sullivan, G.M., Feinn, R., 2012. Using effect size-or why the P value is not enough. J. Grad. Med. Educ. 4, 279–282.
Taylor, L., Schroeder, R., Meyer, E., 2014. Emerging practices and perspectives on Big Data analysis in economics: bigger and better or more of the same? Big Data Soc. 1, 2053951714536877.
Tinati, R., Halford, S., Carr, L., Pope, C., 2014. Big data: methodological challenges and approaches for sociological analysis. Sociology 48, 663–681.
Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, MA.
United Nations, 2007. Register-based Statistics in the Nordic Countries. Review of Best Practices with Focus on Population and Social Statistics, United Nations, New York.
Vittinghoff, E., Glidden, D., Shibosku, S., McCulloch, C., 2005. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. Springer-Verlag, New York.
Wadsworth, M., Kuh, D., Richards, M., Hardy, R., 2006. Cohort profile: the 1946 national birth cohort (MRC national survey of health and development). Int. J. Epidemiol. 35, 49–54.
Webb, P.M., Zimet, G.D., Fortenberry, J.D., Blythe, M.J., 1999. Comparability of a computer-assisted versus written method for collecting health behavior information from adolescent patients. J. Adolesc. Health 24, 383–388.
Winship, C., Morgan, S.L., 1999. The estimation of causal effects from observational data. Annu. Rev. Sociol. 659–706.
Woollard, M., 2014. Administrative data: problems and benefits. A perspective from the United Kingdom. In: Duşa, A., Nelle, D., Stock, G., Wagner, G. (Eds.), Facing the Future: European Research Infrastructures for the Humanities and Social Sciences. SCIVERO, Berlin.
Zhang, L.C., 2012. Topics of statistical theory for register-based statistics and data integration. Stat. Neerl. 66, 41–63.

Please cite this article in press as: Connelly, R., et al., The role of administrative data in the big data revolution in social science research, Social Science Research (2016), http://dx.doi.org/10.1016/j.ssresearch.2016.04.015